Accès au Contenu Informationnel pour les Masses de Données de Documents

Grappa LILLE 3 - UR Futurs INRIA
MOSTRARE

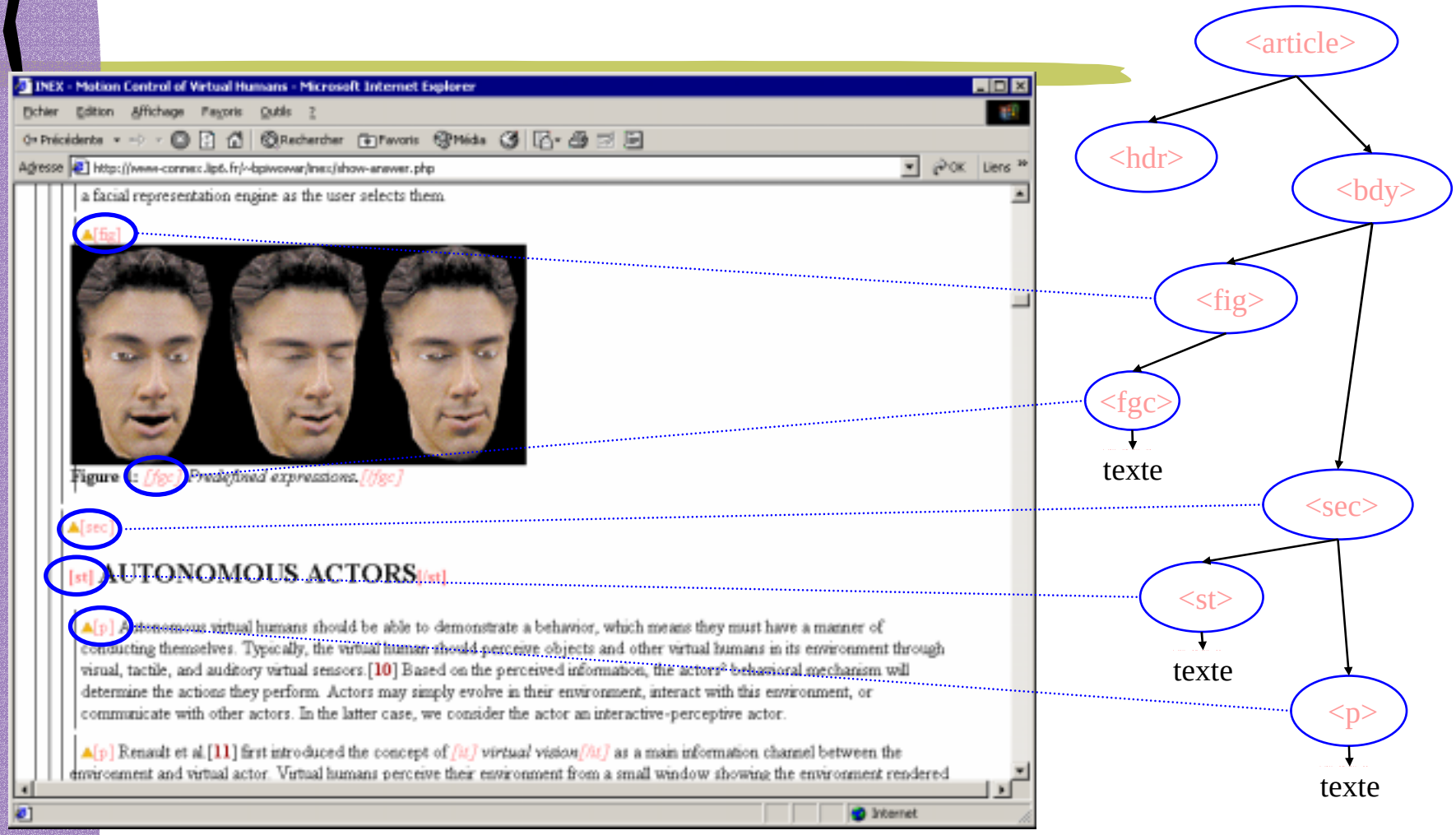
Laboratoire d'Informatique de Paris 6

Laboratoire de Recherche en Informatique
Orsay - UR Futurs INRIA GEMO

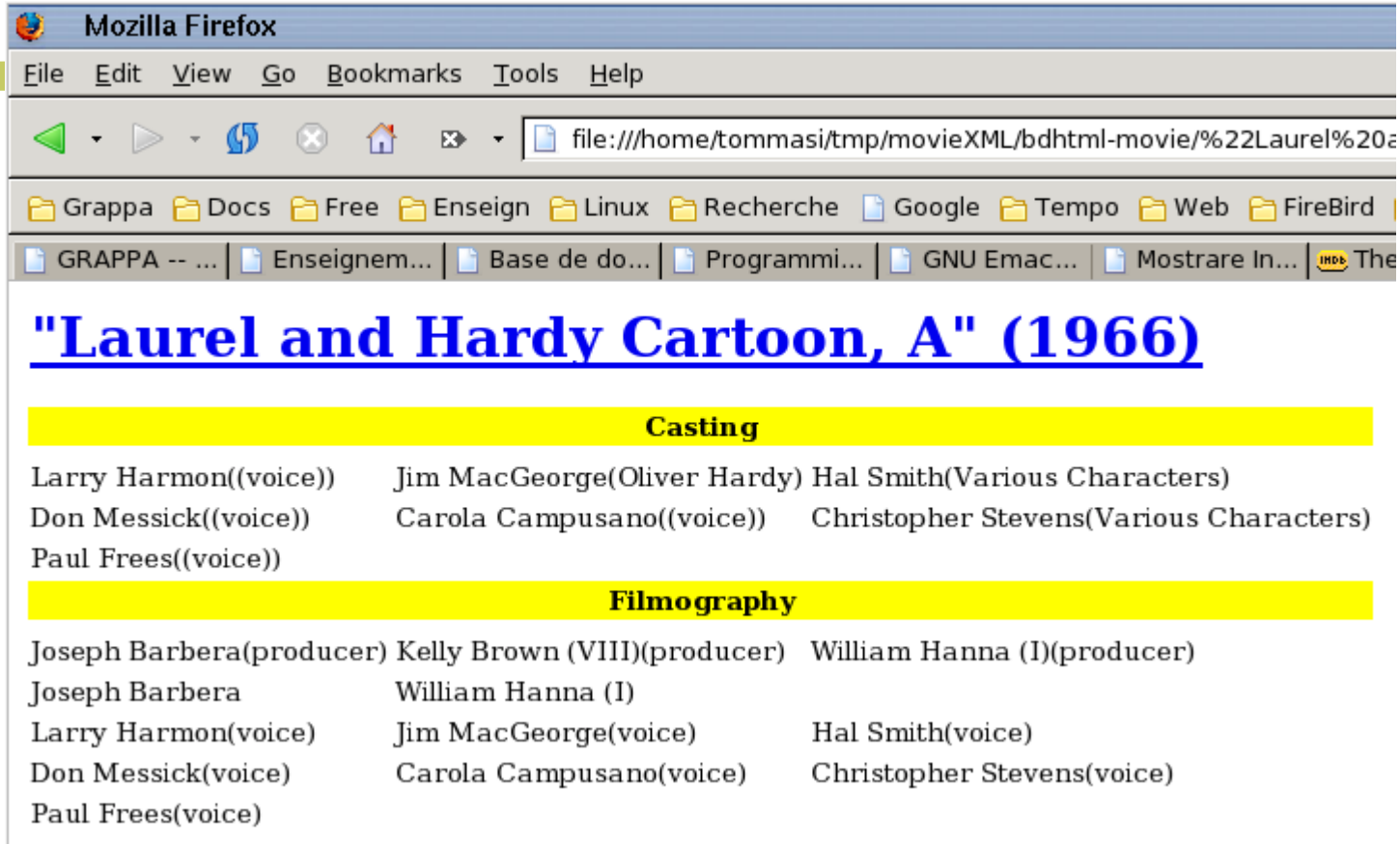
Contexte : Accès à l'information dans des documents semi-structurés

- Changement de la nature des documents
  La structure est de plus en plus présente (documents XML)
  - XHTML (pour les sites Internet)
  - DocBook (pour les documents)
  - Bibliothèques électroniques
- Formats plus riches
  - Contenu
  - Structure logique
  - Sémantique: métadonnées …
  - Multi-média
- Nouveaux besoins
- L'interrogation de ces documents demande des outils capables d'exploiter la richesse des nouveaux formats

XML Documents

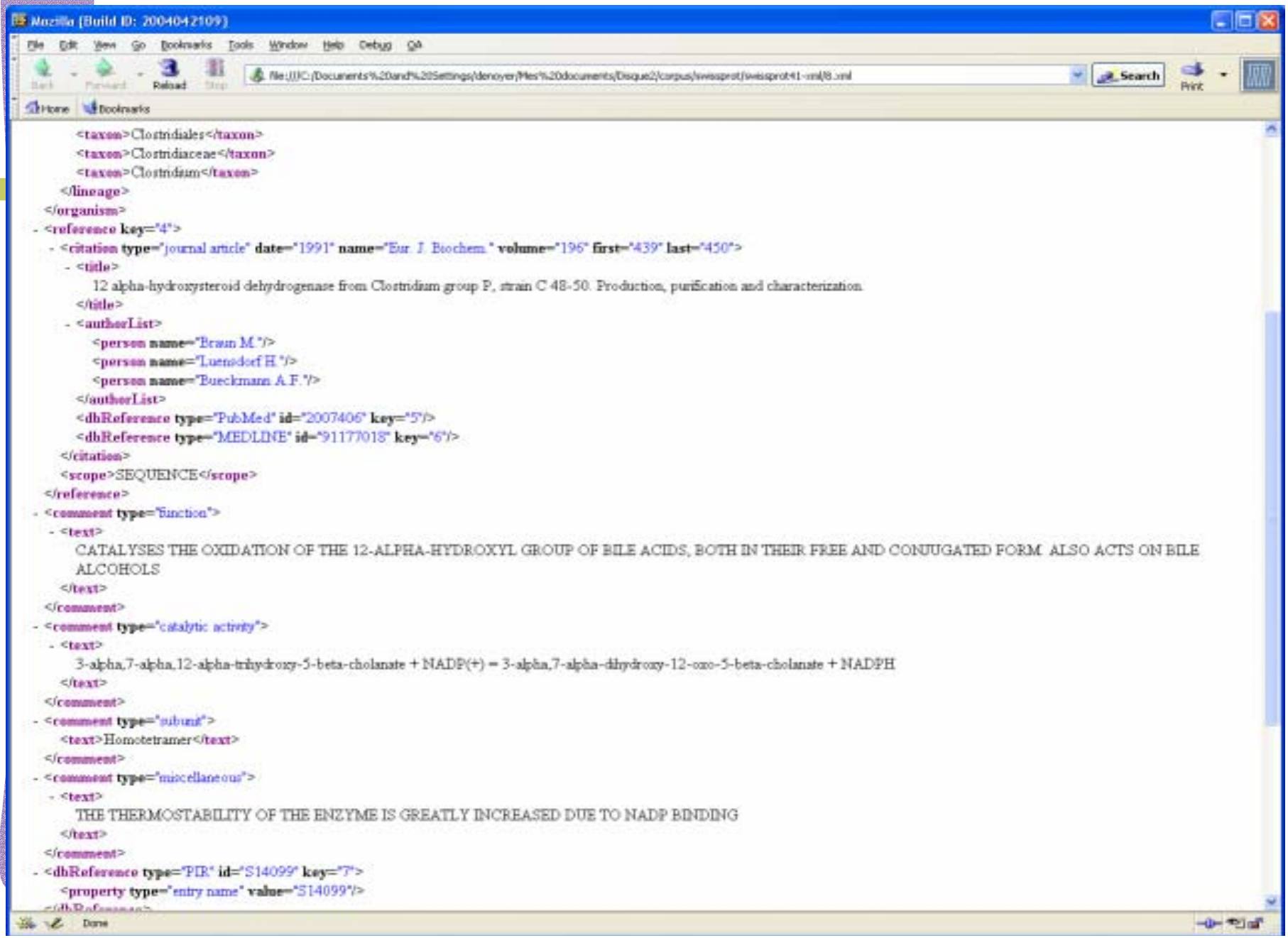


MovieDB


Mozilla Firefox
File Edit View Go Bookmarks Tools Help
file:///home/tommasi/tmp/movieXML/bdhtml-movie/%22Laurel%20a
Grappa Docs Free Enseign Linux Recherche Google Tempo Web FireBird
GRAPPA -- ... Enseignem... Base de do... Programmi... GNU Emac... Mostrare In... The
"Laurel and Hardy Cartoon, A" (1966)
Casting
Larry Harmon((voice)) Jim MacGeorge(Oliver Hardy) Hal Smith(Various Characters)
Don Messick((voice)) Carola Campusano((voice)) Christopher Stevens(Various Characters)
Paul Frees((voice))
Filmography
Joseph Barbera(producer) Kelly Brown (VIII)(producer) William Hanna (I)(producer)
Joseph Barbera William Hanna (I)
Larry Harmon(voice) Jim MacGeorge(voice) Hal Smith(voice)
Don Messick(voice) Carola Campusano(voice) Christopher Stevens(voice)
Paul Frees(voice)

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE movie SYSTEM "movie.dtd" >

<movie>
  <title>"Laurel and Hardy Cartoon, A" (1966)</title>

  <url>http://www.imdb.com/Title?%22Laurel%20and%20Hardy%20Cartoon%
2C%20A%22%20%281966%29</url>
  <aka_titles>
    <aka_title>"Larry Harmon's Laurel &amp;amp; Hardy" (1966)</aka_titl
e>
    <aka_attr>(UK) (video box title)</aka_attr>
  </aka_titles>
  <title_info>
    <tag id="Production_Company">
      <item>Hanna-Barbera Productions [us]</item>
    </tag>
  </title_info>

  <title_info>
    <tag id="Distributor">
      <item>Allworld Telefilm Sales [us]</item>
      <item>Wolper Productions</item>
    </tag>
  </title_info>

  <title_info>
    <tag id="Country_of_Production">
  USA
    </tag>
  </title_info>

  <title_info>
    <tag id="Running_Time">
      <item attr="(156_episodes)">5</item>
    </tag>
  </title_info>

  <title_info>
    <tag id="Filmed_In">
      <item>Color</item>
    </tag>
  </title_info>

  <title_info>
    <tag id="Language">
      <item>English</item>
    </tag>
  </title_info>
```

Swissprot


Mozilla [Build ID: 2004042109]
File Edit View Go Bookmarks Tools Window Help Debug QA
Reload
Search
Print
Home
Bookmarks
<taxon>Clostridiales</taxon>
<taxon>Clostridiaceae</taxon>
<taxon>Clostridium</taxon>
</lineage>
</organism>
<reference key="4">
<citation type="journal article" date="1991" name="Eur. J. Biochem." volume="196" first="439" last="450">
<title>
12 alpha-hydroxysteroid dehydrogenase from Clostridium group P, strain C 48-50. Production, purification and characterization.
</title>
<authorList>
<person name="Braun M."/>
<person name="Luensdorf H."/>
<person name="Bueckmann A.F."/>
</authorList>
<dbReference type="PubMed" id="2007406" key="5"/>
<dbReference type="MEDLINE" id="91177018" key="6"/>
</citation>
<scope>SEQUENCE</scope>
</reference>
<comment type="function">
<text>
CATALYSES THE OXIDATION OF THE 12-ALPHA-HYDROXYL GROUP OF BILE ACIDS, BOTH IN THEIR FREE AND CONJUGATED FORM. ALSO ACTS ON BILE ALCOHOLS
</text>
</comment>
<comment type="catalytic activity">
<text>
3-alpha,7-alpha,12-alpha-trihydroxy-5-beta-cholanate + NADP(+) = 3-alpha,7-alpha-dihydroxy-12-oxo-5-beta-cholanate + NADPH
</text>
</comment>
<comment type="subunit">
<text>Homotetramer</text>
</comment>
<comment type="miscellaneous">
<text>
THE THERMOSTABILITY OF THE ENZYME IS GREATLY INCREASED DUE TO NADP BINDING
</text>
</comment>
<dbReference type="PIR" id="S14099" key="7">
<property type="entry name" value="S14099"/>
Done

Swissprot : 1 entrée par protéine

- Core
  - Séquence
  - ref. bibliographiques
  - taxonomie
- Annotations
  - Fonction(s) de la protéine
  - Structure (secondaire, tertiaire..)
  - Similarité avec autres protéines
  - Maladies liées à des déficiences dans la protéine
  - Liens autres bases .....

Objectifs du projet

- Trois thématiques
  - Recherche d'Information
  - Extraction d'Information
  - Clustering

- Trois familles d'approches issues de l'apprentissage
  - Modèles stochastiques
  - Inférence grammaticale
  - Ensembles fréquents

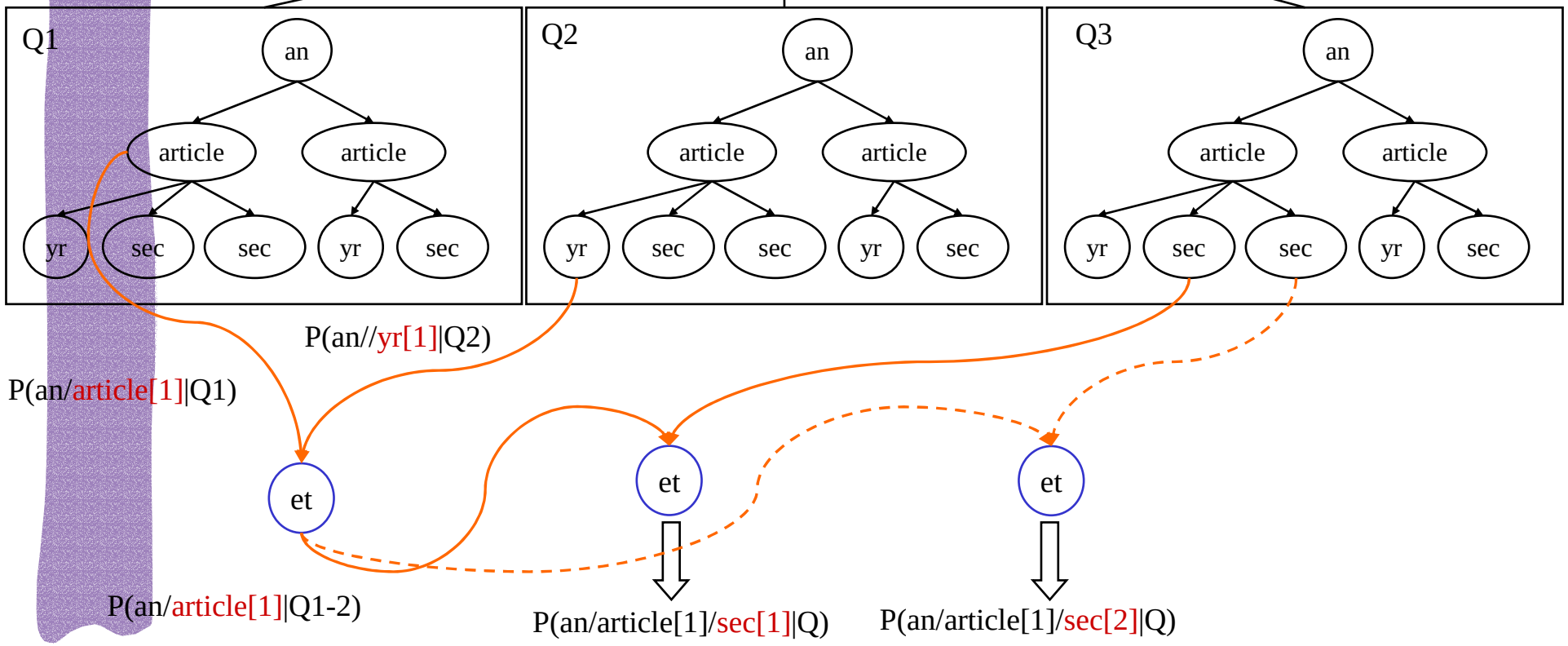
Exemple : RI « CAS » queries

Unité d'information : Doxel

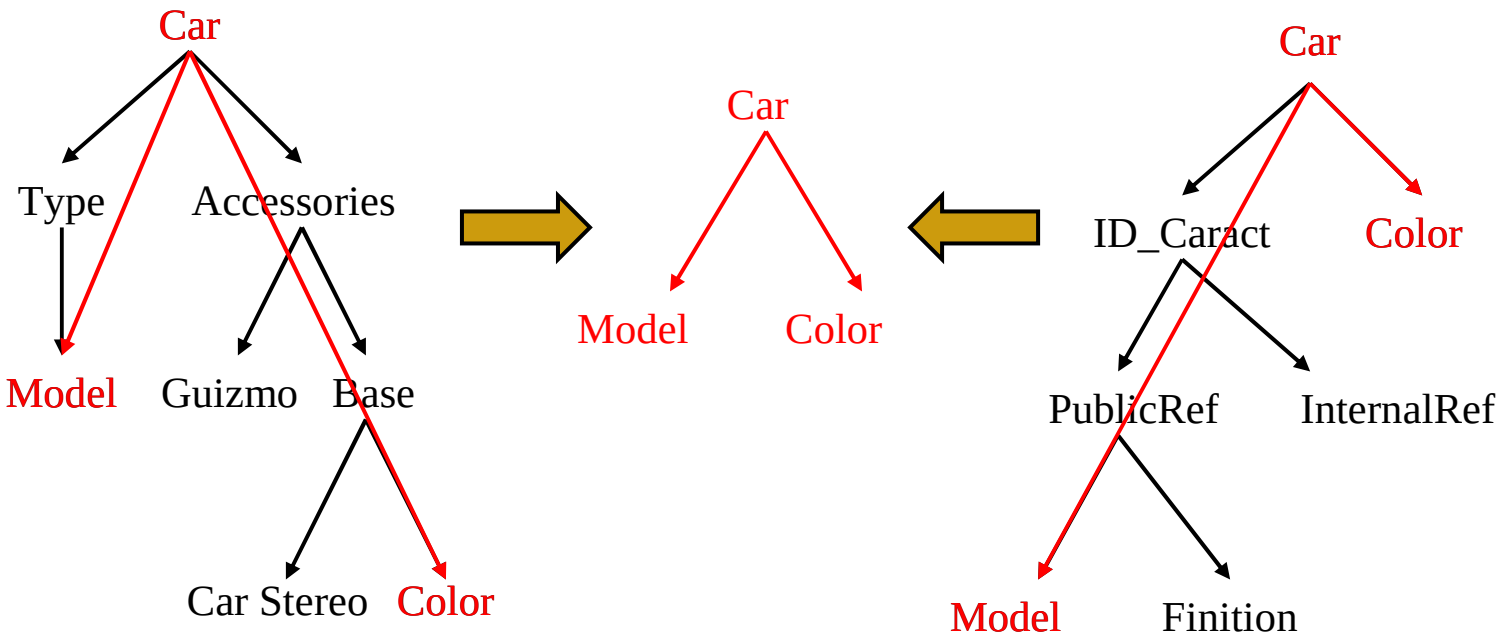
« I want a section on XML in an article about RI published in 2000 »

```
//article[about(., 'RI') and yr>=2000]//sec[about(.,'XML')]
```

Q1 Q2 Q3



Exemple : Clustering Arbres fréquents



T_1 T_2

Travail en cours - méthodes

- But
  - Meilleure définition des problèmes besoins - méthodes
  - Adaptations aux données relationnelles – arbres
  - Adéquation méthode-problématique
  - Complémentarité des approches
  - Liens entre approches

- Travail réalisé : prototypes
  - Moteur de recherche XML (Réseaux Bayesiens)
  - Algorithmes de clustering
    - Structures, structure + contenu
  - Extraction d'information (Grammaires d'arbres et stochastique)

Travail réalisé : données

- Constitutions de bases XML
  - INEX (Journaux IEEE : 16 K articles – 8 M doxels)
  - MovieDB (370 K descriptions films/ émissions TV)
    - Conversion format XML
    - Outils de génération de formats XML hétérogènes
  - Toxgene
    - IBM Toronto lab. / Toronto university: Outil déclaratif pour la génération de corpus de documents XML (corpus / ou bien document unique)
  - Exploitation de bases existantes
    - e.g. Swissprott

Objectifs à 3 ans

- Meilleure définition des problèmes besoins - méthodes
- Adéquation approches – tâches
- Synthèse
- Prototypes
- Ensembles de données
- Participations challenges internationaux
  - INEX (RI) – REX Pascal (Extraction)

Fonctionnement

- Site WIKI
  - http://www.grappa.univ-lille3.fr/twiki/bin/view/Acimdd
  - Documents
  - Bases de données
  - Outils logiciels
- Réunions (5 X 1 journée)

Publications

- A. Termier, M.-C. Rousset, and M. Sebag. Dryade, 2004, :a new approach for discovering closed frequent trees in heterogeneous tree databases. In *International Conference on Data Mining (ICDM)*
- Julien Carme and Joachim Niehren and Marc Tommasi, Querying Unranked Trees with Stepwise Tree Automata, International Conference on Rewriting Techniques and Applications, Lecture Notes in Computer Science 3091, 105 -- 118 4.
- Julien Carme and Aurélien Lemay and Joachim Niehren, 2004, Learning Node Selecting Tree Transducer from Completely Annotated Examples, 7th International Colloquium on Grammatical Inference, Lecture Notes in Artificial Intelligence 8.
- J. Carme and R. Gilleron and A. Lemay and A. Terlutte and M. Tommasi, 2004, Residual Finite Tree Automata, 7th International Conference on Developments in Language Theory, Lecture Notes in Computer Science 2710, 171 -- 182 10.
- F. de Comite and R. Gilleron and M. Tommasi, 2003, Learning Multi-label Alternating Decision Trees from Texts and Data, Proceedings of Intern. Conference on Machine Learning and Data Mining, Lecture Notes in Artificial Intelligence 2734, 35-49 1.
- P. Marty and F. Torre, 2004, Codages et connaissances en extraction d'information, Actes de la Sixième Conférence Apprentissage CAp'2004, 207—222
- Vu (Huyen-Trang), Piwowarski (Benjamin), Gallinari (Patrick), 2004, Filtering in XML Retrieval: a Prospective Analysis
- In XML and Information Retrieval workshop of SIGIR 2004
- Denoyer (Ludovic), Gallinari (Patrick), 2004, Document Structure Matching for heterogeneous corpora In SIGIR 2004
- Piwowarski (Benjamin), Gallinari (Patrick), 2004, An algebra for probabilistic XML Retrieval In The First Twente Data Management Workshop
- Piwowarski (Benjamin), Lalmas (Mounia), 2004, Interface pour l'évaluation de systèmes de recherche sur des documents XML
- In Premiere COnference en Recherche d'Information et Applications (CORIA'04)
- Denoyer (Ludovic), Gallinari (Patrick), 2004, Bayesian Network Model for Semi-Structured Document Classification, In Information Processing and Management